Develop a language-independent extractive text summarization system that generates a summary for a given document

Loading Libraries:

NLTK is a platform for building Python programs to work with human language data. It provides easy-to-use interfaces, text processing libraries for classification, tokenization, stemming, tagging, parsing, and semantic reasoning

```
import nltk
nltk.download('punkt')
nltk.download('stopwords')
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize, sent_tokenize
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Loading and reading file (111.txt) Note: this file is to be uploaded, and can be found in unlabelled datasets provided in the drive link as per lms

```
filename="summarization_dataset/news_articles/001.txt"
f = open("111.txt", "r")#creating a file object
text=f.read() #Read the contents of the file into text
f.close()
```

Printing the content of the file that was just read

```
print(text)
```

```
WHO's Covid weapons fight still $16.8 bn short

The World Health Organization's global appeal for funding for coronavirus v
-- almost half its total needs, the WHO said Tuesday. The funding shortfal
to fight the pandemic, with access to vaccines woefully uneven. WHO chief T
warned that the pandemic remained in a "very dangerous phase" more than 18
```

Pre-processing:

here we convert the text in lower case and remove the stop words. stop words can be imported using nltk packages. A stop word is a commonly used word (such as “the”, “a”, “an”, “in”) that a search engine has been programmed to ignore, both when indexing entries for searching and when retrieving them as the result of a search query.

```
sent_tokens = nltk.sent_tokenize(text.lower())
word_tokens = nltk.word_tokenize(text)
word_tokens_lower=[word.lower() for word in word_tokens]
stopWords = list(set(stopwords.words("english")))
word_tokens_refined=[x for x in word_tokens_lower if x not in stopWords]
print(len(word_tokens_refined))
```

```
75
```

```
print(sent_tokens[0])
print(word_tokens[:8])
print(word_tokens_lower[:8])
print(word_tokens_refined[:8])
```

```
who's covid weapons fight still $16.8 bn short

the world health organization's global appeal for funding for coronavirus v
-- almost half its total needs, the who said tuesday.
['WHO', "'s", 'Covid', 'weapons', 'fight', 'still', '$', '16.8']
['who', "'s", 'covid', 'weapons', 'fight', 'still', '$', '16.8']
["'s", 'covid', 'weapons', 'fight', 'still', '$', '16.8', 'bn']
```

Here we simply count the number of times a word appears in the document, and thus formulate the Frequency Distribution.

```
freqTable = dict()
for word in word_tokens_refined:
    if word in freqTable:
        freqTable[word] += 1
    else:
        freqTable[word] = 1
print(len(freqTable))
```

```
58
```

Taking reference from the frequencytable we Compute score of each sentence. This score will be used as a basis to include or not include the statement in the summary.

```
sentenceValue = dict()
for sentence in sent_tokens:
    for word in nltk.word_tokenize(sentence):
        if word in freqTable.keys():
            if sentence in sentenceValue:
                sentenceValue[sentence] += freqTable[word]
            else:
                sentenceValue[sentence] = freqTable[word]
```

Now that the score for each is computed, we will find an average score.

If the score is satisfacctory we add it to our summary sentence.

Finally we can print this sentence and which is the final summary.

```
sumValues = 0
for sentence in sentenceValue:
    sumValues += sentenceValue[sentence]
average = int(sumValues / len(sentenceValue))
print(average)
# Storing sentences into our summary.
summary = ''
for sentence in sent_tokens:
    if (sentence in sentenceValue) and (sentenceValue[sentence] > (1.1*average)):
        summary += " " + sentence
print(summary)
```

```
43
 who's covid weapons fight still $16.8 bn short

the world health organization's global appeal for funding for coronavirus 
-- almost half its total needs, the who said tuesday.
```

✓ 0s completed at 6:31 PM